Naiveté and sophistication in dynamic inconsistency

Mathematical Social Sciences 87 (2017) 40–54

Contents lists available at ScienceDirect

Mathematical Social Sciences

journal homepage: www.elsevier.com/locate/econbase

Naiveté and sophistication in dynamic inconsistency

Zsombor Z. Méder [a,*], János Flesch [b], Ronald Peeters [c]

[a] Humanities, Arts, and Social Sciences, Singapore University of Technology and Design, Singapore
[b] Department of Quantitative Economics, Maastricht University, The Netherlands
[c] Department of Economics, Maastricht University, The Netherlands

HIGHLIGHTS

- We contrast naiveté and sophistication by separating intentions and beliefs.
- Intentions and beliefs of a decision maker's various selves are analyzed.
- Hybrid decision makers who are sometimes naive, sometimes sophisticated can be modeled.

ARTICLE INFO

Article history:
Received 4 September 2015
Received in revised form
14 October 2016
Accepted 3 February 2017
Available online 11 February 2017

ABSTRACT

This paper introduces a general framework for dealing with dynamic inconsistency in the context of Markov decision problems. It decouples and examines concepts that are often entwined in the literature. It distinguishes between the decision maker and her various temporal selves, and between the beliefs and intentions of the selves. The creation of a unified formalism to deal with dynamic inconsistency allows for the introduction of a hybrid decision maker, who is naive sometimes, sophisticated at others. Such a hybrid decision maker can be used to model situations where type determination is endogenous. Interestingly, the analysis of hybrid types indicates that self-deception can be optimal.

1. Introduction

Imagine that you are sitting with friends, drinking beer. You have just finished your second pint, and your friends want to order another round. You think to yourself: 'well, I could deal with one or two more, but then I really should go home'. However, you are also acutely aware from previous experience that after your third pint your mindset is likely to change: you will start fooling yourself, repeating over and over in the course of the evening: 'just one more beer, and then I am *really* going home'. This would lead to an undesirable outcome, getting drunk and having a hangover the next morning. So you wisely leave your friends after just your second beer. What is happening here? The framework that we propose here makes it possible to model this and similar scenarios.[1]

Traditionally, there are three main ways to portray decision makers under time inconsistency. The first one regards decision makers as naifs (Akerlof, 1991; O'Donoghue and Rabin, 1999b), the second attributes sophistication to them (Laibson, 1997; Fischer, 1999; Harris and Laibson, 2001), while the third argues that even resolute behavior is possible (McClennen, 1990). A common assumption of these models in the classic papers on dynamic inconsistency is regarding the decision maker as falling entirely into one of the above three categories, treating her type as exogenously given. More recently, mainly building on the work of O'Donoghue and Rabin (2001), hybrid decision makers have been considered, in models of so-called 'partial naiveté'. However, these models still treat this type as exogenous to the decision problem.

One way to interpret our example above is that you are sophisticated after finishing the first two beers, but as you drink more, you expect to become naive later on. The story indicates that in certain situations, a decision maker could cause her type to change; moreover, she might even be able to reason about such changes. Any perspective that assumes a fixed type is unable to capture such situations. In order to fix this shortcoming, we attempt a general interpretation of naiveté and sophistication for dynamic inconsistency. The language we develop allows the introduction of hybrid-type decision makers such as the one in our example.

Following the review of the relevant literature, we build a formalism that allows for precise definitions of the two

* Corresponding author.
E-mail addresses: mederzsombor@gmail.com (Z.Z. Méder), j.flesch@maastrichtuniversity.nl (J. Flesch), r.peeters@maastrichtuniversity.nl (R. Peeters).

[1] The model of this situation is presented in Section 7.

http://dx.doi.org/10.1016/j.mathsocsci.2017.02.002
0165-4896/

most commonly discussed types of decision makers, naifs and sophisticates. We work in discrete time, assuming that the situation of the decision maker can be captured as a Markov decision problem. We distinguish between the level of the decision maker and her multiple selves. Starting on the level of the selves, we specify both the intentions and the beliefs of each self. Next, the properties of these intention–belief pairs (coherence, stationarity, consistency) are discussed. Moving to the level of the decision maker (i.e., the collection of all selves), we define the concept of a *frame*, and consider its properties. We link the properties of intention–belief pairs to the properties of frames. The concept of a frame is novel. A frame provides all the relevant information about a decision maker facing dynamic inconsistency.

After clarifying our assumptions on utility functions, we define and introduce the two types of decision making, naiveté and sophistication. We provide existence results for optimal frames of both types, and discuss various properties of such frames. In the main text body, we then introduce decision makers with a hybrid type, also providing an existence theorem, and discuss two examples of hybrid decision making in detail. In the concluding section, we point towards further extensions of the model.

The contributions of this paper are thus threefold. First, it provides new concepts and distinctions for problems of dynamic inconsistency. In particular, the representation of the decision maker through a frame, and the theorems relating consistency and stationarity could prove useful (Section 3.6). Second, it provides definitions for naiveté and sophistication, and shows the existence of naively and sophisticatedly optimal frames. Third, it introduces hybrid naive-sophisticated types, expanding the scope of dynamic inconsistency models. As a corollary, the analysis of hybrid types shows that self-deception can be optimal.

2. Related literature

Modeling approaches to dynamic inconsistency come in two varieties (Asheim, 2007). Dual-self planner–doer models bear a close analogy to principal–agent models (Thaler and Shefrin, 1981). A (single) *planner*, endowed with dynamically consistent preferences formulates plans; a present-biased *doer* can execute them or deviate from them. The conceptual background of planner–doer models is multifold: They sometimes rely on the hot–cold empathy gap (Loewenstein, 2005), on recent findings of neuroscience, or even the Freudian distinction between the *id* and the *ego*. Fudenberg and Levine (2006) argue in favor of a dual-self model as being analytically simpler, more in line with findings in neuroscience, and nevertheless being able to explain a large number of empirical phenomena. A key advantage of this approach is that welfare comparisons are relatively straightforward: The preferences of the planner are generally adopted to be normatively relevant. Recent models allow for the planner to learn about the doer's type through costly experimentation (Ali, 2011), or can include self-control (Gul and Pesendorfer, 2001; Benabou and Pycia, 2002; Fudenberg and Levine, 2012).

An important limitation of many dual-self models is that they assume that individuals have long-run, time-consistent preferences (e.g., Benabou and Pycia, 2002; Fudenberg and Levine, 2006; Brocas and Carrillo, 2008). In this paper we avoid this assumption, and instead adopt the multiple-self approach.[2] Our primary focus is on naive and sophisticated decision makers, and hybrid types. Multiple-self models have analyzed naive and sophisticated decision makers in a variety of settings. For instance, Laibson (1994), Fischer (1999), Laibson (1997) and Angeletos et al. (2001) work with sophisticates, while Akerlof (1991) and O'Donoghue and Rabin (1999b) assume naiveté. In a general equilibrium setting, Herings and Rohde (2006) and Herings and Rohde (2008) deal with both naifs and sophisticates. O'Donoghue and Rabin (1999a) is a cardinal paper, as it compares and contrasts naiveté and sophistication for the case of quasi-hyperbolic discounting, with so-called immediate costs and rewards.

The first model of hybrid types can be found in O'Donoghue and Rabin (2001). They use quasi-hyperbolic ($\beta - \delta$) discounting to define partial naiveté. While naifs think their β is 1, the actual β is fully known to sophisticates, while partially naifs think they have a β that is larger than their actual one. Partially naifs thus entertain false beliefs about the future (just like naifs). This approach has proven its fruitfulness especially in the contract design literature (DellaVigna and Malmendier, 2004; Eliaz and Spiegler, 2006; Gilpatric, 2008); while DellaVigna and Malmendier (2006) focus on a monopolistic firm facing a mixed population of consumers. Heidhues and Kőszegi (2009) work with partial naifs who are 'weakly optimistic' regarding their future present-biasedness.

A significant result of the O'Donoghue and Rabin (2001) approach is that – compared to sophistication – *any* degree of partial naiveté can generate arbitrarily large losses in efficiency for the decision maker. In this sense, the limit of partial naiveté (as the perceived present-biasedness approaches the real parameter) is not sophistication.

There exist a few other attempts to analyze hybrid decision making in the literature. In Asheim (2007), selves have a perceived preference persistence, i.e., a probability (between 0 and 1) with which they think their preferences will be identical in the next period. However, this belief is always incorrect, as their preferences will change with probability 1. In Jehiel and Lilico (2010), selves endowed with exponential discounting have access to information about a number of future periods ('foresight'). They find that improving the length of foresight always improves welfare.

Hybrid decision making in our paper bears a resemblance to the model of Bernheim and Rangel (2004). They analyze addictive behavior by distinguishing between a 'cold' (sophisticated) and a 'hot' decision making mode (where preferences and choice may diverge). In the cold mode, the decision maker is able to reason about what her future behavior will be in either mode. This is similar in spirit to how we treat sophisticated selves in hybrid decision making in Section 7. The main difference between their work and ours is in the treatment of naiveté. For Bernheim and Rangel (2004), use of the addictive substance in the hot mode is not a matter of deliberation and choice, but is instead a 'mistake' triggered by environmental cues. Thus, being in the hot mode already determines the behavior of the decision maker, and no proper decision making takes place. In contrast, our naive selves may have a number of choices available to them, and they are able to reason about these choices and the future. Furthermore, in the current work, naiveté is not necessarily disadvantageous for the decision maker, and so the behavior it leads to need not be evaluated as mistaken.

This paper expands on the existing literature by providing the foundations of a hybrid model that is independent of the quasi-hyperbolic assumption of O'Donoghue and Rabin (2001), and where type determination is an endogenous part of the decision problem, as in our motivating example.

3. Basic concepts

In this section, we introduce our framework and notations. Standard definitions are provided for the notions of 'decision problem' and 'history'. We then proceed by defining 'intentions' and 'beliefs' on the level of the selves, as well as 'frame', on the level of the decision maker. Towards the end of the section, we present some results on the relationship between consistency and stationarity.

[2] Bach and Heilmann (2011) link multiple-self models to the philosophical literature on personal identity.

3.1. Markov decision problem

We start with a decision maker facing a finite Markov decision problem on an infinite horizon.[3]

Definition 1. A *finite Markov decision problem* is given by:

- the set of time periods $T = \{0, 1, 2, \ldots\}$;
- a finite set of states Ω, with $\omega_0 \in \Omega$ as the initial state;
- a finite and nonempty set of pure actions A_ω that the decision maker can choose from in state ω;
- a payoff function $u_\omega : A_\omega \to \mathbb{R}$ that assigns a payoff to every action in state ω;
- transition probabilities $m_\omega : A_\omega \to \Delta(\Omega)$, with $m_\omega\left(\omega'|a_\omega\right)$ denoting the probability to transit from state ω to state ω' when action a_ω is chosen.

Note that this definition excludes the possibility of randomization over actions.

3.2. History

To capture all the informational aspects on which the choice of an action can be conditioned, we introduce the notion of a history:

Definition 2. A *history* h has the form $h = \left(\omega_0, a_{\omega_0}, \ldots, \omega_{t-1}, a_{\omega_{t-1}}, \omega_t\right)$, with:

- $\omega_i \in \Omega$, for $i \in \{0, 1, \ldots, t\}$;
- $a_{\omega_i} \in A_{\omega_i}$, for $i \in \{0, 1, \ldots, t-1\}$;
- $m_{\omega_i}\left(\omega_{i+1}|a_{\omega_i}\right) > 0$, for $i \in \{0, 1, \ldots, t-1\}$.

The current time at h is denoted by $t = t(h)$, and the function $\omega(h) = \omega_{t(h)}$ indicates the current state at history h. We use H to refer to the set of all histories.

If history h' begins with h, we say that h' *succeeds* h, and denote this with $h' \rhd h$. The subset of H that consists of all histories that succeed h is denoted by $H^{\rhd h}$:

$H^{\rhd h} = \{h' \in H | h' \rhd h\}$.

We refer to $h_0 = (\omega_0)$ as the 'root history'. To shorten notation, when specifying a history, we sometimes omit the commas separating states and actions, and also the enclosing parentheses. Thus, history $h = \left(\omega_0, a_{\omega_0}, \omega_1\right)$ will be occasionally written as $h = \omega_0 a_{\omega_0} \omega_1$.

3.3. Conceptual foundations

The fundamental entities in our model are *selves*. A self is associated with a particular history—we will thus speak of a 'self at history h'. We emphasize that for two fundamental reasons, it is not sufficient to talk merely about a self at a particular time period. First, depending on past decisions, either the assets, information, and circumstances pertaining to the decision, or the preferences, or even the reasoning capacities of the self might vary. Second, reasoning about one's own future behavior often requires the consideration of off-equilibrium, counterfactual behavior. Consider, for instance, someone reasoning about the hot–cold empathy gap (Loewenstein, 2005), and a choice that can either lead to a 'hot', or a 'cold' state. Suppose one contemplates the action that will lead to the 'cold' state. Whether this choice is optimal will depend on the behavior of the future self in the 'hot' state, which has different circumstances, preferences, and more limited reasoning capacities than the self in the 'cold' state. Thus, in the context of our paper, it would be inadequate to talk about a 'self at time period t'. Instead, when talking about a self, we will refer to the history at which that particular self exists.

From a temporal perspective, selves relate both the past, the present, and the future. Selves are identified by the *past* sequence of states and actions. They form beliefs and intentions about the *future*. Finally, they have the ability to choose and execute an action in the *present*, based on their preferences, beliefs and intentions.[4]

We keep the general assumption that past actions have no effect on the well-being generated by current and future actions of the selves, i.e., 'bygones are bygones'. Each self has full control over her current actions.[5]

Using the distinction between experienced utility and decision utility (Kahneman et al., 1997), we can delimit two senses of 'expected utility from taking action a'. Let us disregard the immediate payoff for taking the action, and consider only future payoffs. In one sense, the phrase could mean 'experienced utility from expectation', i.e., utility that *is actually experienced* by a particular self due to expecting a certain stream of future payoffs. Think of a student that decides to study for an upcoming exam instead of watching her favorite TV show. She might, in fact, *already* enjoy the benefits of the decision to study (she is already less anxious about the exam, maybe she relishes the idea that she is doing 'the right thing', etc.). The other sense of 'expected utility' could be rendered as 'the expected present value of various streams of payoffs' that the self's current decision can lead to. In this sense, the self does not experience any actual change in utility by choosing one or other course of action; she is merely able to calculate with these future payoffs. This distinction between the two senses of expected utility will be used in Section 4, and in our conceptual interpretations of naiveté and sophistication in subsequent sections.

3.4. Intentions and beliefs

We aim to give a full description for the two most prevalent decision maker types (naifs and sophisticates) and the hybrid types that we introduce later. For this purpose, we deal with three components: the current action, the intended future actions, and the belief about what future selves will in fact do. There is no special reason for assuming that the latter two coincide for future actions, although with our definitions, they coincide for naifs and sophisticates, but not for hybrid decision makers. To simplify notation, we reduce this triadic framework to just intentions and beliefs, and assume that for the current action, these two have to coincide: No self can be wrong about which action she takes, and each self takes the action that she intends *at that moment*.

We emphasize that dealing with both intentions and beliefs is not standard. In fact, the most common practice is to conflate the two concepts. Think of backward induction, for instance. When reasoning about another player, the picture is clear: player 1 forms beliefs on what player 2 might do, given that they will choose optimally. However, what about player 1's future behavior? When player 1 is rational, expects to stay so, and expects to have no

[3] The latter is not a restrictive requirement, since it is easy to rewrite a decision problem on a finite horizon to one on an infinite horizon.

[4] In summarizing the literature on self-control and self-management, Cowen (1991) identifies a self "with a set of preferences linked to certain cognitive and volitional capacities". We conceptualize the self along the same lines: In Section 3.4, we deal with cognitive and volitional capacities, while in Section 4, we discuss preferences. To refer to the collection of *all* selves, we use the notion of the *decision maker*, and we introduce that level of analysis in Section 3.6.

[5] To appreciate that this choice is not so obvious, see Jehiel and Lilico (2010). Similarly, Elster interprets the Ulysses story in such a way that control over the current action is essentially eliminated for anyone listening to the sirens (Elster, 1979).

change in preferences, player 1 *both wants and expects* to choose optimally at future nodes, that is, her intentions and beliefs match. However, when player 1's preferences change between the present and the future, this coherence between beliefs and intentions can break down: she might want to behave in a certain way in the future, but cannot reasonably expect to actually do so. Suppose, for instance, player 1 contemplates whether to visit the pub to drink two pints in the evening. She knows that her preferences will change after the second pint, and she will want to stay for more. She intends to drink two pints in the evening, but does not believe she will actually do so.[6]

The basic building blocks of our model are all functions from the set of histories that succeed the present to the set of available actions at those histories.

Definition 3. The *intentions* of a self at history $\bar{h}$ assign an intended action to each history that succeeds the present:

$$i^{\bar{h}} : h \in H^{\triangleright \bar{h}} \mapsto A_{\omega(h)}.$$

Definition 4. The *beliefs* of a self at history $\bar{h}$ assign an action to each history that succeeds the present:

$$b^{\bar{h}} : h \in H^{\triangleright \bar{h}} \mapsto A_{\omega(h)}.$$

Note that intentions and beliefs are defined at all succeeding histories, even at those that are not intended to or believed to be reached by the particular self.

We now proceed to define an intention–belief pair for a particular self.

Definition 5. An *intention–belief* pair at history $\bar{h}$ is a pair of intentions and beliefs for that self, with the added property that the belief and intention for the current action coincide:

$$s^{\bar{h}} = \left(i^{\bar{h}}, b^{\bar{h}}\right), \quad \text{with } i^{\bar{h}}\left(\bar{h}\right) = b^{\bar{h}}\left(\bar{h}\right).$$

The set of all intention–belief pairs for this self is denoted by $S^{\bar{h}}$.

For a self at $\bar{h}$, $i^{\bar{h}}(h)$ refers to the intention, while $b^{\bar{h}}(h)$ refers to the belief component of the intention–belief pair $s^{\bar{h}}$ regarding history h. For example, $b^{\bar{h}}(h) = a$ should be read as: 'the self at $\bar{h}$ believes the self at h will choose action a'.

We can now define stationarity, coherence, as well as consistency for intention–belief pairs.

Definition 6. The intentions (or beliefs) of a self at $\bar{h}$ are *stationary* whenever the intended (believed) actions depend only on the end-state. Thus, $i^{\bar{h}}$ or $b^{\bar{h}}$ is called stationary if, for all $h, h' \in H^{\triangleright \bar{h}}$ with $\omega(h) = \omega\left(h'\right)$, we have $i^{\bar{h}}(h) = i^{\bar{h}}\left(h'\right)$ or respectively, $b^{\bar{h}}(h) = b^{\bar{h}}\left(h'\right)$. An intention–belief pair $s^{\bar{h}}$ is stationary if both its constituent intentions $i^{\bar{h}}$ and beliefs $b^{\bar{h}}$ are stationary.

For example, if each day of the week is modeled as a single state, the intentions of a self who intends to eat in a restaurant every second Saturday, and to stay home on every other one, is not stationary.

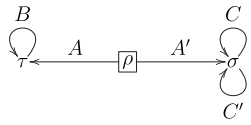

Fig. 1. Stationary intention–belief pairs with intransitive consistency.

Definition 7. A self at $\bar{h}$ is said to hold a *coherent* intention–belief pair, if her intentions and beliefs about future actions coincide for all future histories. Formally, an intention–belief pair $s^{\bar{h}} = (i^{\bar{h}}, b^{\bar{h}})$ of a self at $\bar{h}$ is coherent if $i^{\bar{h}}(h) = b^{\bar{h}}(h)$ for all $h \in H^{\triangleright \bar{h}}$.

For example, the intention–belief pair of a self who intends to stop drinking, but believes she will be unable to do so, is not coherent.

Definition 8. The intention–belief pairs of two selves at h and h' are said to be *consistent* if they assign the same intentions and beliefs to each history that succeeds both selves, i.e., s^h and $s^{h'}$ are consistent, if $s^h\left(h''\right) = s^{h'}\left(h''\right)$ for all $h'' \in H^{\triangleright h} \cap H^{\triangleright h'}$.[7]

For example, an intention–belief pair formulated yesterday which intended eating apples today and an intention–belief pair formulated today which intends eating cookies instead are not consistent.

While coherence concerns the relationship between the intentions and beliefs of the same intention–belief pair, i.e., belonging to one self, consistency compares intention–belief pairs of two distinct selves.

A natural question is whether the consistency of intention–belief pairs is transitive, i.e., whether the consistency of s^h and $s^{h'}$ and the consistency of $s^{h'}$ and $s^{h''}$ imply that s^h and $s^{h''}$ are also consistent. If $h'' \triangleright h' \triangleright h$, then this is indeed the case. However, without this constraint, consistency is not transitive in general—it is not even transitive within the set of stationary intention–belief pairs. To see this, take the decision problem in Fig. 1.[8] We construct three stationary intention–belief pairs s^{ρ}, $s^{\rho A \tau}$, and $s^{\rho A' \sigma}$ such that s^{ρ} and $s^{\rho A \tau}$ are consistent, as well as $s^{\rho A \tau}$ and $s^{\rho A' \sigma}$, but s^{ρ} and $s^{\rho A' \sigma}$ are not consistent. Also, let $s^{\rho}(\rho) = (A, A)$, $s^{\rho}(h) = (B, B)$ if $\omega(h) = \tau$, and $s^{\rho}(h) = (C, C)$ if $\omega(h) = \sigma$. Intuitively, s^{ρ} means: 'I choose A, believe and intend B in state τ, and believe and intend C in state σ'. Define two other intention–belief pairs through $s^{\rho A \tau}(h) = (B, B)$ for all $h \triangleright (\rho, A, \tau)$ ('do B after you reach τ'), and $s^{\rho A' \sigma}(h) = (C', C')$ for all $h \triangleright (\rho, A', \sigma)$ ('do C' after you reach σ'). All of these intention–belief pairs are stationary. Clearly, s^{ρ} and $s^{\rho A \tau}$ are consistent, since they both require the decision maker to choose B in state τ, and after history (ρ, A, τ) no state other than τ is reachable. Next, $s^{\rho A \tau}$ and $s^{\rho A' \sigma}$ are consistent, since histories (ρ, A, τ) and (ρ, A', σ) neither succeed, nor precede each other. But s^{ρ} and $s^{\rho A' \sigma}$ are not consistent, as they assign different actions to the state σ. This shows that consistency of intention–belief pairs is not transitive on the set of stationary intention–belief pairs.

6 Another illustration is provided by the so-called 'toxin puzzle' (Kavka, 1983). The original formulation is as follows: "An eccentric billionaire places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed. All you have to do is… intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin". In this example, I intend to intend to drink the toxin, but, most likely, I do not believe that I will intend to drink it.

7 So, if two intention–belief pairs are defined at histories that neither succeed nor precede each other, then they are consistent, as there are no histories that succeed both.

8 We use figures like this one to represent decision problems. States are denoted by Greek characters; in this decision problem, we have states ρ, τ and σ, and $\boxed{\rho}$ indicates ρ to be the initial state. For actions, we use Roman capitals—in this case, we have action A, A' (available in state ρ), B (the only action available in τ), and C, C' (in σ). Our examples involve only deterministic transition probabilities, and the transitions associated with each state are represented by arrows. For example, for the decision problem in Fig. 1 choosing action C in state σ leads to the state σ with probability 1.

3.5. Truncation

The following definitions of 'truncation', albeit technical in nature, are necessary for the definition of a stationary frame. Truncation formalizes the idea of 'bygones are bygones', and chips away an initial segment of the history.

Definition 9. Take any history $h = (\omega_0, a_{\omega_0}, \ldots, \omega_{t(h)})$. The *truncation* operator $\dashv_k$, defined for any $k \leq t(h)$, removes the first k pairs of this sequence, so $\dashv_k h = (\omega_k, a_{\omega_k}, \ldots, \omega_{t(h)})$. For a set of histories $H' \subseteq H$, we refer to the set of k-truncated histories by $\dashv_k H'$.

Similarly, a truncated intention–belief pair is defined for a self who 'forgot' (or disregards) all of her past; future histories obviously do not include descriptions of forgotten segments of the past anymore. Thus, the truncated intention–belief pair of a self at $\bar{h}$ will be defined on the set $\dashv_{t(\bar{h})} H^{\triangleright \bar{h}}$.

Definition 10. For any intention–belief pair $s^{\bar{h}}$, the *truncated* intention–belief pair $\dashv s^{\bar{h}} : h \in \dashv_{t(\bar{h})} H^{\triangleright \bar{h}} \mapsto A_{\omega(h)}$ denotes the function for which $\dashv s^{\bar{h}} \left(\dashv_{t(\bar{h})} h\right) = s^{\bar{h}}(h)$, for all $h \in H^{\triangleright \bar{h}}$.

We note that while truncated histories are defined for truncations of arbitrary length ($\dashv_k$), for intention–belief pairs we only need truncations of length $t(\bar{h})$ for a self at history $\bar{h}$.

To see this definition at work, think of considering to stop smoking on the first day of the next month. Take a self who resolves on July 24th: 'I intend to, and will stop smoking from August 1st' and then fails. On August 24th, she forms another intention–belief pair: 'I intend to, and will stop smoking from September 1st'. It is easy to see that these intention–belief pairs are not consistent: For instance, they prescribe different smoking behavior for August 28th—the first intention–belief pair is incompatible with it, while the second allows it. However, there is an intuitive sense in which they are very similar. Indeed, they map them into the same resolve that uses indexicals instead of precise dates: 'I may smoke for one more week, and then I intend to and will stop'.[9] Truncating the present history highlights this similarity by getting rid of the past. In our example, the original intention–belief pairs are not identical or consistent; but their truncated versions are identical.

3.6. Frames

We now move from the level of the selves to the level of the decision maker. Since there is no *a priori* reason for the selves to have consistent intention–belief pairs, different selves can form different intentions and entertain different beliefs about any certain future self. To have an 'external' overview of all selves, we introduce the concept of a frame. In our terminology, a frame is an auxiliary tool for representing the intention–belief pairs of all possible selves. In this way, a frame contains a full description of the intentions and beliefs under all contingencies, i.e., at all histories.[10]

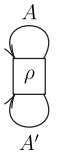

Fig. 2. A basic decision problem.

Table 1
An example of a frame for the decision problem in Fig. 2.

		$h \in H^{\triangleright \bar{h}}$					
		ρ	$\rho A \rho$	$\rho A' \rho$	$\rho A \rho A \rho$	$\rho A \rho A' \rho$	$\cdots$
$\bar{h}$	ρ	$A'A'$	AA'	AA'	AA	$A'A'$	$\cdots$
	$\rho A \rho$	–	$A'A'$	–	AA'	$A'A$	$\cdots$
	$\rho A' \rho$	–	–	AA	–	–	$\cdots$
	$\rho A \rho A \rho$	–	–	–	$A'A'$	–	$\cdots$
	$\rho A \rho A' \rho$	–	–	–	–	AA	$\cdots$
	$\cdots$	–	–	–	–	–	$\cdots$

Definition 11. A *frame* assigns an intention–belief pair to each self. That is, a frame is a function $f : h \in H \mapsto S^h$.

Fig. 2 shows an extremely simple decision problem, for which an example of a frame is represented in Table 1. Each entry is a pair of As and A's, an intended action and a belief about an action. Each row corresponds to an intention–belief pair for a self at $\bar{h}$, defining an intention and a belief for each history that succeeds $\bar{h}$. For example, the entry AA' for row $\bar{h} = (\rho A \rho)$ and column $h = (\rho A \rho A \rho)$ should be interpreted as such: The self at $(\rho A \rho)$ intends to choose action A at history $(\rho A \rho A \rho)$, while believing the self at $(\rho A \rho A \rho)$ will, in fact, choose action A'. The frame thus specifies the intentions and beliefs of all selves over all other (present and future) selves. Our definition of an intention–belief pair ensures that on the diagonal of the table, the intentions and the beliefs match.

We now proceed to introduce three properties of frames. Our definition of stationarity makes use of the truncation operator defined above.

Definition 12. A frame f is said to be *stationary*, if only the end-state matters when assigning intention–belief pairs to histories, i.e., for any histories h and h', if $\omega(h) = \omega\left(h'\right)$, then $\dashv f(h) = \dashv f(h')$.

Stationarity of a frame is different from the stationarity of the intention–belief pairs involved. For the decision problem in Fig. 2, Table 2 offers an example of a non-stationary frame of stationary intention–belief pairs. To check this, what needs to be verified first is that each row represents a stationary intention–belief pair. As there is only one state for this decision problem, this means that in each row, we should see the same intention–belief pair, which is indeed the case. Thus, Table 2 shows a frame of stationary intention–belief pairs. The frame itself however is not stationary: By truncating the intention–belief pairs in the first and second row, we get a different intention–belief pair.

On the other hand, Table 3 shows a stationary frame of non-stationary intention–belief pairs. It is easy to see that it is a frame of non-stationary intention–belief pairs, as each row represents one intention–belief pair, which are not stationary—for instance,

[9] Actually, the difference between specifying a future consumption period by a *calendar date* or through its *temporal distance* from the present has already been noticed by Strotz (1956). This difference is experimentally explored by Read et al. (2005), finding that subjects only exhibit hyperbolic discounting when future periods are identified via their temporal distance.

[10] It should be noted that the term 'frame' is already used in psychology and behavioral economics, in a different sense. Tversky and Kahneman (1981) use the term 'decision frame' "to refer to the decision maker's conception of the acts, outcomes, and contingencies associated with a particular choice". To paraphrase them, we could say we use the term 'frame of a decision maker' to refer to "the analyst's conception of the acts, outcomes, and contingencies (i.e., intentions, beliefs, and actions) associated with a particular decision maker". In the psychological approach to decision making, several frames can be associated with a single decision problem, whereas we assume that an analyst will have a single frame of a decision maker, the frame describes the decision maker's multiple selves correctly. There are, of course, many other important differences between the two meanings of the term. However, such proliferation of meanings does not necessarily lead to confusion, as long as the context is clear.

Table 2
Non-stationary frame of stationary intention–belief pairs.

		$h \in H^{\triangleright \bar{h}}$					
		ρ	$\rho A \rho$	$\rho A' \rho$	$\rho A \rho A \rho$	$\rho A \rho A' \rho$	$\cdots$
$\bar{h}$	ρ	AA	AA	AA	AA	AA	AA
	$\rho A \rho$	–	$A'A'$	–	$A'A'$	$A'A'$	$A'A'$
	$\rho A' \rho$	–	–	AA	–	–	AA
	$\rho A \rho A \rho$	–	–	–	AA	–	AA
	$\rho A \rho A' \rho$	–	–	–	–	$A'A'$	$A'A'$
	$\cdots$	–	–	–	–	–	$\cdots$

Table 3
Stationary frame of non-stationary intention–belief pairs.

		$h \in H^{\triangleright \bar{h}}$					
		ρ	$\rho A \rho$	$\rho A' \rho$	$\rho A \rho A \rho$	$\rho A \rho A' \rho$	$\cdots$
$\bar{h}$	ρ	$A'A'$	AA	AA	AA	AA	AA
	$\rho A \rho$	–	$A'A'$	–	AA	AA	AA
	$\rho A' \rho$	–	–	$A'A'$	–	–	AA
	$\rho A \rho A \rho$	–	–	–	$A'A'$	–	AA
	$\rho A \rho A' \rho$	–	–	–	–	$A'A'$	AA
	$\cdots$	–	–	–	–	–	$\cdots$

$s^{\rho}(\rho) = (A', A') \neq (A, A) = s^{\rho}(\rho A \rho)$. The frame itself is stationary, which can be checked by comparing the rows. As there is only one state, we need to compare the truncations of all intention–belief pairs. It can be read from Table 3 that by 'forgetting the past', we get the same intention–belief pair in each row, namely: The self picks action A' right away, intends to choose action A, and believes her future selves will always do so in the future.

For a real-life example of a stationary frame of non-stationary intention–belief pairs, think of the decision maker who, waking up every day, decides to take *just one more* shot of heroin, and intends (as well as believes) to quit the next day.

Next, we define a consistent frame. The intuitive idea is that a frame is consistent if no deviation can be expected from previous intentions and beliefs.

Definition 13. A frame f is said to be *consistent* if the intention–belief pairs $f(h)$ and $f(h')$ assigned to any two histories h and h' are consistent.

Consistency is a very strong notion. A consistent decision maker – at whichever history she is contemplating the present and the future – would never change her mind about any intention–belief pair. An example would be a heroin user who goes cold turkey immediately and definitely, never ever restarting her substance use. If she relapses, she will not be consistent anymore: Her choice to quit for good implies that at the time of quitting her intentions and beliefs for the future history at which she relapses do not match her intentions and beliefs at the point of relapse for the (then-)current history.

According to this definition, if f is a consistent frame, then we get $f(h)\left(h''\right) = f\left(h'\right)\left(h''\right)$ whenever $h'' \in H^{\triangleright h} \cap H^{\triangleright h'}$, and either $h' \in H^{\triangleright h}$, or $h \in H^{\triangleright h'}$. Note that, since for a consistent frame, $f(h)(h) = f(h')(h)$ for all h' and $h \in H^{\triangleright h'}$; and our requirement for intention–belief pairs that $i^h(h) = b^h(h)$, a consistent frame is necessarily made up by coherent intention–belief pairs. This implies that a choice of an action for all histories uniquely determines a consistent frame. Similarly, a choice of an action for all states uniquely determines a consistent frame of stationary intention–belief pairs.

Theorem 1. *A consistent, stationary frame consists of stationary intention–belief pairs.*

Proof. Take any histories h, h' and h'' with $h' \triangleright h$ and $h'' \triangleright h$ for which $\omega(h') = \omega(h'')$. We have to show that $f(h)\left(h'\right) = f(h)\left(h''\right)$.

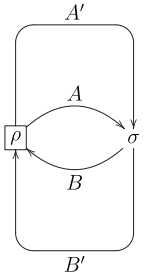

Fig. 3. Stationary frame of stationary intention–belief pairs is not necessarily consistent.

For this, see that:

$$f(h)\left(h'\right) = f\left(h'\right)\left(h'\right) = \dashv f\left(h'\right)\left(\dashv_{t(h')} h'\right)$$
$$= \dashv f\left(h''\right)\left(\dashv_{t(h'')} h''\right) = f\left(h''\right)\left(h''\right) = f(h)\left(h''\right).$$

For the respective equations, we use, in order, consistency, definition of truncation, stationarity of the frame, definition of truncation, and consistency again. □

Theorem 2. *A consistent frame of stationary intention–belief pairs is a stationary frame.*

Proof. Take histories h, h', h'', h''' with $\omega(h) = \omega\left(h'\right)$. We have to show that

$$(\dashv f(h))\,(\dashv_{t(h)} h'') = (\dashv f(h'))\,(\dashv_{t(h')} h''')$$

if $\dashv_{t(h)} h'' = \dashv_{t(h')} h'''$. Note that such h'' and h''' exist, because the end-states in h and h' are identical. Using the fact that $\omega(h'') = \omega(\dashv_{t(h)} h'') = \omega(\dashv_{t(h')} h''') = \omega(h''')$, we get:

$$(\dashv f(h))\,(\dashv_{t(h)} h'') = f(h)(h'') = f(h_0)(h'') = f(h_0)(h''')$$
$$= f(h')(h''') = (\dashv f(h'))\,(\dashv_{t(h')} h''').$$

We use, in turn, the definitions of truncation, consistency, stationarity of the intention–belief pair, consistency, and finally, truncation again. (The root history h_0 is one which is surely succeeded by both h'' and h'''.) □

Based on the two theorems above, one might expect that a stationary frame of stationary intention–belief pairs would be consistent. However, this is not necessarily so, as can be seen from the following example. Consider the Markov decision problem in Fig. 3. For all $h \triangleright \bar{h}$, let:

$$s_1^{\bar{h}}(h) = \begin{cases} (A, A) & \text{if } \omega(h) = \rho, \\ (B, B) & \text{if } \omega(h) = \sigma, \end{cases} \quad \text{and}$$

$$s_2^{\bar{h}}(h) = \begin{cases} (A', A') & \text{if } \omega(h) = \rho, \\ (B', B') & \text{if } \omega(h) = \sigma. \end{cases}$$

Now, let us define a frame f, so that:

$$f(\bar{h}) = \begin{cases} s_1^{\bar{h}} & \text{if } \omega(\bar{h}) = \rho; \\ s_2^{\bar{h}} & \text{if } \omega(\bar{h}) = \sigma. \end{cases}$$

This is obviously a frame of stationary intention–belief pairs. It also is a stationary frame, since only the end-state matters in assigning an intention–belief pair to a history, according to the definition. However, it is not a consistent frame, since:

$$f(\rho)(\rho A \sigma) = s_1^{\rho}(\rho A \sigma) = (B, B) \neq (B', B')$$
$$= s_2^{\rho A \sigma}(\rho A \sigma) = f(\rho A \sigma)(\rho A \sigma).$$

3.7. Induced intention–belief pair

We assume that, since each self has control over her current action (and only that), the actual actions executed by each self h can be obtained from frame f by looking at $f(h)(h)$, in other words, from the diagonal of the frame.

Definition 14. The *induced intention–belief pair* of a frame f specifies the actual actions chosen by each self:

$$\Lambda(f) : h \in H \mapsto A_{\omega(h)}, \quad \text{given by } \Lambda(f)(h) = f(h)(h).$$

The induced intention–belief pair of the frame represented in Table 1 is $\Lambda(f)(\rho) = (BB)$, $\Lambda(f)(\rho A \rho) = (BB)$, $\Lambda(f)(\rho B \rho) = (AA)$ etc.

It is handy to define for some frame f, and a self at $\bar{h}$, the induced intention–belief pair for the (present and) future:

$$\Lambda^{\triangleright\bar{h}}(f) : h \in H^{\triangleright\bar{h}} \mapsto A_{\omega(h)}, \quad \text{given by } \Lambda^{\triangleright\bar{h}}(f)(h) = f(h)(h);$$
$$\Lambda^{\blacktriangleright\bar{h}}(f) : h \in H^{\triangleright\bar{h}} \setminus \{\bar{h}\} \mapsto A_{\omega(h)}, \quad \text{given by}$$
$$\Lambda^{\blacktriangleright\bar{h}}(f)(h) = f(h)(h).$$

3.8. Remarks

The considerations of this section highlight the interdependencies between various concepts. Coherence reflects a match between intentions and beliefs, desires and reality. In addition, coherence is a necessary condition for the consistency of frames (see the discussion immediately preceding Theorem 1). Consistency of a frame ensures that selves are not 'let down' by future selves, in the sense that expected behavior matches actual future behavior. In various contexts, stationarity of intentions, beliefs, and stationarity of a frame can also be desirable properties. Primarily, stationary intentions "prescribe the simplest form of behavior consistent with rationality" (Maskin and Tirole, 2001). Finally, a stationary frame describes a decision maker who is stable over time.

The language developed here can be helpful for discussing problems of dynamic inconsistency formulated in the multiple-self framework, and is independent of the specific assumptions on utility functions of the next section. In particular, we believe that our distinction between beliefs/intentions, selves/decision maker, intention–belief pairs/frame, as well as our method of representing frames through tables should prove useful.

4. Utility and discounting

The term *payoff*, introduced in Definition 1, refers to the immediate gains or losses resulting from an action. Formally, a payoff gained in period t is denoted by u_t, and a stream of payoffs starting at period t by $u_{t\rightarrow} = (u_t, u_{t+1}, \ldots, o)$. We say that a stream of payoffs $u_{t\rightarrow}$ starting at period t coincides with a stream of payoffs $u'_{t'\rightarrow}$ starting at period t' if $u_t = u'_{t'}$ and $u_{t+1} = u'_{t'+1}$, and so on.

Payoffs are fully determined by the decision problem, the state and the action taken. However, time preference implies that identical payoffs might be regarded differently by various selves. Throughout the paper, we make two assumptions on the utility functions $U^h(u)$ that integrate a stream of future payoffs into a single number. The first assumption states that U^h is continuous at infinity for every h, and is adapted from Fudenberg and Levine (1983).

Assumption 1. U^h is continuous at infinity for every h, i.e., for any ϵ, there is a horizon $Q(h)$ – possibly depending on h – such that the total variation of utility after $t(h) + Q(h)$ is less than ϵ:

$$\sup_{u,u'} |U^h(u) - U^h(u')| < \epsilon.$$

Another crucial assumption is that selves are identical in the way they evaluate streams of payoffs, i.e., we assume stationary preferences (Peleg and Yaari, 1973).

Assumption 2. For any two selves at histories h and h', and coinciding streams of payoffs $u_{t(h)\rightarrow}$, $u'_{t(h')\rightarrow}$, the utilities of the two selves are equal, i.e., $U^h(u_{t(h)\rightarrow}) = U^{h'}(u'_{t(h')\rightarrow})$.

If the utility functions satisfy First and Second Order Separability[11] (Lapied and Renault, 2012), then the discount factor for a future payoff u_t can only depend on the time distance $t - t(h)$. In our examples, we use a discounted utility function of a particular form, namely, quasi-hyperbolic discounting:

$$U^h(u_{t(h)\rightarrow}) = u_{t(h)} + \beta \sum_{t=t(h)+1}^{\infty} \delta^{t-t(h)} u_t,$$

with $0 \leq \beta \leq 1$ and $0 \leq \delta < 1$.

In Section 3.3, we distinguished between two senses of the term 'expected utility': utility *actually experienced* from expecting a future payoff stream, and 'expected utility' as simply *a means of calculating* with various future courses of action. This distinction is formally nailed down and further refined by the following definitions.[12]

Definition 15. The *expected utility based on intentions* of the intention–belief pair $s^h = (i^h, b^h)$ for a self at h is:

$$U_i^h\left(s^h\right) = \mathbb{E}\left[i^h\right]\left(U^h\right).$$

The next definition focuses on how much utility a self can reasonably expect *ex ante*, whenever making utility calculations:

Definition 16. The *expected utility based on beliefs* of an intention–belief pair $s^h = (i^h, b^h)$ for a self at h is:

$$U_b^h\left(s^h\right) = \mathbb{E}\left[b^h\right]\left(U^h\right).$$

Our final definition disregards mere expectations, and captures the utility gained by a self considering which actions future selves will have *actually* implemented under various eventualities. Using the notion of an induced intention–belief pair, we can define *induced utility*:

Definition 17. Given a frame f, the (ex post) *induced utility* of the root self at h_0 is:

$$U_r(f) = \mathbb{E}\left[\Lambda^{\triangleright h_0}(f)\right]\left(U^{h_0}\right).$$

Notice that traditionally, the above three meanings of the term 'expected utility' coincide. The reason is that whenever dynamic inconsistency does not pose a problem, intentions and beliefs regarding future actions coincide; moreover, the decision maker always executes the intentions of past selves.

5. Naiveté

At first sight, it is not even clear whether naiveté is a property of the decision maker or that of a self. In this and the following

[11] First Order Separability means that preferences over a set of outcomes are such that there is no interaction between the outcomes of various periods. Second Order Separability allows the isolation of effects of temporal distance.

[12] Sáez-Martí and Weibull (2005) connect discounting with altruism towards future selves. They note that "[c]urrent welfare or 'total utility', so defined, does not stem only from current instantaneous utility but also from (the anticipation of) the stream of future instantaneous utilities". Our distinction highlights exactly this: Is it the actual future stream, or the anticipation of that stream that is really important?

section, we first define naiveté and sophistication for selves. A decision maker will be defined as naive (or sophisticated) whenever all her multiple selves are naive (or sophisticated). We introduce hybrid types, and return to our original example presented in Section 1 after analyzing these base cases.

Naiveté has been characterized in several ways in the literature; a naif is aware or unaware of different things, depending on the particular interpretation. A naif is said to:

- choose at each stage an option which seems currently the best (Strotz, 1956; Hammond, 1976);
- fail to realize that future selves will have different preferences (O'Donoghue and Rabin, 1999a, 2001; Sarafidis, 2004; DellaVigna and Malmendier, 2006; Herings and Rohde, 2006; Heidhues and Kőszegi, 2009);
- believe that – though her preferences might change – she has perfect self-control about the future, allowing her to commit (O'Donoghue and Rabin, 1999a, 2001; Gruber and Kőszegi, 2001; Ali, 2011).

It is easy to see that these are genuinely alternative interpretations of naiveté. A common aspect is that something is amiss with the beliefs held by the selves. We argue that these troubles arise from the way the naif determines her beliefs. In particular, for a naive self, her current preferences determine her intentions, which in turn determine her beliefs on future actions. Thus, it does not matter whether a self holds an explicit belief on the lack of change in her preferences, or whether she believes she will simply fail to act on such changes, or that she has strong beliefs in her own will- or pre-commitment power. The essential feature of naiveté is the sequence of determination presented in Fig. 4.[13]

Definition 18. An intention–belief pair s^h of a self at h is *naively optimal*, if it maximizes expected utility based on intentions:

$$s^h \in \operatorname*{argmax}_{s \in S^h} U_i^h(s),$$

and if it is coherent:

$$b^h\left(h'\right) = i^h\left(h'\right), \quad \text{for all } h' \in H^{\triangleright h}.$$

A frame f is naively optimal if the intention–belief pair $f(h)$ is naively optimal at each history h.

Working in a continuous-time discounted utility framework, Strotz (1956) shows that only when the discount function is exponential does the decision maker possess a consistent naively optimal frame for all decision problems. For any non-exponential discount functions there are decision problems for which there is no consistent naively optimal frame. However, the existence of a stationary naively optimal frame is guaranteed under our assumptions.

Theorem 3. *For any decision problem, there exists a stationary naively optimal frame.*

Proof. For each h, the set $\operatorname{argmax}_{s \in S^h} U_i^h(s)$ is nonempty, because the set S^h is nonempty and compact,[14] and U_i^h is continuous (Assumption 1). Therefore, the set of intention–belief pairs where the maximum is in fact reached is nonempty. But note that the optimality condition in the definition of naively optimal intention–belief pair only determines the intention component, and thus, beliefs can be constructed freely. This means that we can ensure coherence, i.e., we can choose a naively optimal intention–belief pair at each h.

13 To connect to our discussion in Section 3.3, we can regard the naive as someone who maximizes expected utility based on intentions; she behaves as if future expected utility could already be experienced, irrespective of what will actually happen in the future.

14 See Fudenberg and Levine (1983), p. 261–262 for more details.

Preferences → intentions → beliefs → intention-belief pair.

Fig. 4. The forming of intentions and beliefs by a naive self.

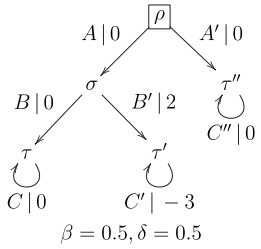

Fig. 5. Multiplicity of induced utility for naively optimal frames.

Now, to guarantee that the generated frame is stationary, we need to choose the same truncated intention–belief pair for each set of histories where the end-state is identical. This is always possible, since whenever the final state is identical for two histories, both the set of truncated intention–belief pairs and the utility function defined at those histories are identical (Assumption 2), and therefore so are the set of truncated optimal intention–belief pairs. □

Is there a unique stationary naively optimal frame? For a naive decision maker, this depends on whether for each history, there is a unique naively optimal intention–belief pair; this latter problem can be reduced to whether for each *state* that can be reached, there is a unique naively optimal intention–belief pair (defined at any history where the current state is that state). For generic decision problems, it seems likely that this is indeed the case, though the scope of the proof is beyond this work. Now, if there is a unique naively optimal intention–belief pair for each history, then there is only one naively optimal frame—and it is stationary, too. In degenerate cases, where multiple naively optimal intention–belief pairs can be assigned to at least one state, we get stationary naively optimal frames, along with non-stationary ones. It should also be noted that – because of the possibility of inconsistency – multiplicity of naively optimal frames also leads to a multiplicity of induced utilities.

As an example, consider the decision problem on Fig. 5.[15] Discounting is quasi-hyperbolic with $\beta = \delta = 0.5$. Since naively optimal intention–belief pairs are by definition coherent, we will focus on such intention–belief pairs. There are only two histories where the action choice is not trivial, $h_0 = (\rho)$ and $h_1 = (\rho, A, \sigma)$, so the naive root self at $h_0 = (\rho)$ has to consider only four coherent intention–belief pairs, which we will denote, by $s^{h_0}_{AB}$, $s^{h_0}_{AB'}$, $s^{h_0}_{A'B}$ and $s^{h_0}_{A'B'}$. They are defined as:

1. $s^{h_0}_{AB}(h_0) = (A, A)$ and $s^{h_0}_{AB}(h_1) = (B, B)$;
2. $s^{h_0}_{AB'}(h_0) = (A, A)$ and $s^{h_0}_{AB'}(h_1) = (B', B')$;
3. $s^{h_0}_{A'B}(h_0) = (A', A')$ and $s^{h_0}_{A'B}(h_1) = (B, B)$;
4. $s^{h_0}_{A'B'}(h_0) = (A', A')$ and $s^{h_0}_{A'B'}(h_1) = (B', B')$.

It can be easily seen that from the perspective of the root self, $U_i^{h_0}(s^{h_0}_{AB}) = U_i^{h_0}(s^{h_0}_{A'B}) = U_i^{h_0}(s^{h_0}_{A'B'}) = 0 > -\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}(2 - \frac{1}{2} \cdot \frac{3}{\frac{1}{2}}) = U_i^{h_0}(s^{h_0}_{AB'})$. Thus, the naive root self is indifferent between choosing action A first, and B afterwards, or simply A'.

Now we focus on the self at h_1. There are two possible intention–belief pairs, which we will denote by $s^{h_1}_{B}$ and $s^{h_1}_{B'}$. They are determined by:

15 Whenever payoffs are explicit in a figure, the payoff associated with an action is written in the form $A|0$, meaning that choosing action A generates a payoff of 0.

1. $s_B^{h_1}(h_1) = (B, B)$;
2. $s_{B'}^{h_1}(h_1) = (B', B')$.

For the self at h_1, $U_i^{h_1}(s_{B'}^{h_1}) = 2 - \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{\frac{1}{2}} = 0.5 > 0 = U_i^{h_1}(s_B^{h_1})$. Therefore, the naive self at h_1 prefers to choose action B'. Now, consider the two naively optimal frames $f_{A'BB'}$ and $f_{ABB'}$, defined as:

- $f_{A'BB'}(h_0) = s_{A'B}^{h_0}$ and $f_{A'BB'}(h_1) = s_{B'}^{h_1}$;
- $f_{ABB'}(h_0) = s_{AB}^{h_0}$ and $f_{ABB'}(h_1) = s_{B'}^{h_1}$.

From the previous calculations, we see that both $f_{A'BB'}$ and $f_{ABB'}$ are naively optimal frames. However, the induced utilities are not equal: $U_r(f_{A'BB'}) = 0$ and $U_r(f_{ABB'}) = -\frac{1}{4}$. The underlying reason is as follows: The root self at h_0 expects to earn 0 both by having the intention–belief pair $s_{A'B}^{h_0}$ or by $s_{AB}^{h_0}$. But with $s_{AB}^{h_0}$, after getting to history h_1 – on account of being present-biased – she will not stick to her previous intention–belief pair, which would prescribe her choosing action B; instead, she chooses B', which leads to a decrease in her induced utility.

To finalize our discussion of naiveté: If we interpret an optimal frame as predictive for a naive decision maker's behavior, then for some decision problems, we get a unique prediction of actions, intentions and beliefs for each history; and we expect stationary behavior in realization, and a single prediction for the induced utility. On the other hand, in some cases, we get multiple predictions of actions, intentions and beliefs for some selves; we do not necessarily expect stationary behavior; and we do not necessarily get a unique expectation for induced utility.

6. Sophistication

Similarly to naiveté, there are several definitions of sophistication:

- optimality under a credibility constraint: following a feasible optimal plan, or a plan that the decision maker will actually follow (Strotz, 1956; Yaari, 1978);
- game-theoretic notion: an intra-personal subgame-perfect equilibrium, sometimes also referred to as 'Strotz–Pollak equilibrium' (Peleg and Yaari, 1973; Kocherlakota, 1996; Vieille and Weibull, 2009);
- rational expectations: perfectly anticipating future behavior (O'Donoghue and Rabin, 2001; Gilpatric, 2008);
- self-awareness: being aware of future changes in discount rates and preferences (Hammond, 1976; Heidhues and Kőszegi, 2009; Ali, 2011).

These definitions do not match precisely. For instance, the notion of sophistication as an intra-personal equilibrium does not guarantee the satisfaction of rational expectations in case when multiple such equilibria exist. In our notation, it is possible that sophisticated selves at h and h' with $h' \triangleright h$ each have an intention–belief pair s^h, $s^{h'}$ supported by an intra-personal equilibrium, but the action $b^h(h') \neq b^{h'}(h')$; that is, sophistication in itself is no guarantee for consistency. Just like for naiveté, we would like to offer a new interpretation of sophistication. We regard sophistication as primarily a characteristic of selves. In our framework, the defining feature of sophisticated selves is that they first consider their beliefs about the future, and only then do they form intentions (see Fig. 6).[16]

Beliefs and preferences → intentions → action → intention-belief pair.

Fig. 6. The forming of intentions and beliefs by a sophisticated self.

For now, our sophisticated selves assume that all future selves will be sophisticated, too—we will relax this assumption for hybrid selves in Section 7.

A sophisticatedly optimal intention–belief pair is made up, first, by beliefs about future selves. A sophisticated self believes each future self will pick a best response to future selves' choices. Those choices will, of course, depend on the beliefs of the respective future selves. So we implicitly have to consider second-order beliefs, the beliefs of each self about the beliefs of selves about the future. To simplify matters, we assume that the second-order beliefs of a sophisticated self coincide with her first-order beliefs.[17] Thus, the current self believes that future selves will believe what she currently believes.

The second component of a sophisticatedly optimal intention–belief pair to consider is the intentions. These are set to match the beliefs. As the sophisticated self knows she has no control over her future selves, she can just as well intend future actions that future selves are choosing anyway.

Finally, the current action is chosen to be a best response to future actions.

Definition 19. An intention–belief pair s^h of a self at h is *sophisticatedly optimal*, if:

$$b^h\left(h'\right) \in \underset{a \in A_\omega(h')}{\operatorname{argmax}} U_b^{h'}\left(s^h\left[a : h'\right]\right), \quad \text{for all } h' \in H^{\triangleright h},$$

and if it is coherent:

$$i^h\left(h'\right) = b^h\left(h'\right), \quad \text{for all } h' \in H^{\triangleright h},$$

where $s^h\left[a : h'\right]$ denotes the intention–belief pair where the action taken at h' – i.e., the intention and belief for history h' – is replaced with action a in intention–belief pair s^h.[18]

A frame f is sophisticatedly optimal if the intention–belief pair $f(h)$ is sophisticatedly optimal at each history h.

One remark about the intention component of a sophisticatedly optimal intention–belief pairs is in order. It could be argued that instead of intending to give a best response to future beliefs at h' based on the preferences at h', the self at h should intend something else at h'; namely, to give a best response to the choices of future selves based on the preferences at h, not the ones at h'. However, we want to guarantee the coherence of sophisticatedly optimal intention–belief pairs to ensure that a consistent sophisticated intention–belief pair always exists. Moreover, it is psychologically plausible that the sophisticated self wants to maintain this coherence—indeed, this is why she is reasoning about future selves, realizing that the preferences of future selves might be different from her current ones. We will see in Section 7 that for hybrid selves intentions and beliefs might not match.

Theorem 4. *For any decision problem, there exists a consistent sophisticatedly optimal frame.*

16 A sophisticated self can thus be interpreted as someone who maximizes expected based on beliefs—on the particular belief that every future self will reason exactly as she does. She behaves as if future expected utility would merely be *decision* utility, and that her actual expected utility *in the present* would depend on what will, in fact, happen in the future (cf. 3.3).

17 O'Donoghue and Rabin (2001) make the assumption on second-order beliefs explicit for partially naive selves.

18 When calculating $U_b^{h'}(s^h)$, only the payoffs generated by s^h for the histories succeeding h' should be taken into account; i.e., we consider the expected utility induced by s^h for the subtree starting at h'.

Proof. Fix an enumeration of all histories, that is, a bijection $e : \mathbb{N} \to H$. Although it is not necessary for the proof, we can assume that the root history is taken first, then all histories at stage 1 are enumerated, then all histories at stage 2, and so on. Let $\mathscr{A} = \times_{h\in H} A_{\omega(h)}$, where the product is taken in the order according to e. So, an element $a = \left(a^h\right)_{h\in H}$ of $\mathscr{A}$ prescribes, for every history $h \in H$, an action a^h for the self at h. Take an arbitrary $\check{a} \in \mathscr{A}$.

Now, for every $n \in \mathbb{N}$, we construct an $a_n \in \mathscr{A}$ as follows. For every history h beyond stage n, that is, with $t(h) > n$, let $a_n^h = \check{a}^h$. Then, we proceed by backwards induction. For every history h at stage n, that is $t(h) = n$, let a_n^h be an action for the self at h that maximizes her utility if all selves that succeed her play the action according to a_n (or equivalently, according to $\check{a}$). In general for $k \in \{1, \ldots, n\}$, if we have defined a_n^h for all histories h with $t(h) > k$, then for every history h with $t(h) = k$, we choose a_n^h to be an action for the self at h that maximizes her utility if all selves that succeed her play the action according to a_n.

So, we obtain a sequence $(a_n)_{n\in\mathbb{N}}$ in the space $\mathscr{A}$. Note that $\mathscr{A}$ with the product topology is compact, and because H is countable, it is metrizable too. Consequently, it is sequentially compact, which implies that the sequence $(a_n)_{n\in\mathbb{N}}$ has a subsequence $\left(a_{n_k}\right)_{k\in\mathbb{N}}$ which converges to some $\hat{a} \in \mathscr{A}$. This means that, for every $m \in \mathbb{N}$, there exists a $K_m \in \mathbb{N}$ such that, for every $k \geq K_m$, the actions $a_{n_k}^h$ and $\hat{a}^h$ coincide for all histories h with $t(h) \leq m$.

Let s_n^h be the intention–belief pair of the self at history h where the action $a_n^{h'}$ is intended for all histories $h' \rhd h$, and this self believes that the same actions will be chosen—that is, a coherent intention–belief pair. Similarly, we define the intention–belief pair s^h with respect to $\hat{a}$. Now consider the frame f that assigns intention–belief pair s^h to the self at h, for every history h.

First, we show that f is consistent. Take any h, h', the corresponding intention–belief pairs s^h and $s^{h'}$ and some history h'' with $h'' \rhd h$ and $h'' \rhd h'$. Then

$$s^h\left(h''\right) = \left(\hat{a}^{h''}, \hat{a}^{h''}\right) = s^{h'}\left(h''\right).$$

Thus, f is a consistent frame.

We now prove that f is sophisticatedly optimal. For this purpose, consider an arbitrary self, say at history h, and a self at a history $h' \rhd h$. By construction, for every $n \geq t\left(h'\right)$, the action $a_n^{h'}$ maximizes the utility of the self at h' if all selves that succeed her will play the action according to a_n. Thus,

$$b_n^h\left(h'\right) = a_n^{h'} \in \underset{a\in A_{\omega\left(h'\right)}}{\operatorname{argmax}} U_b^{h'}\left(s_n^h\left[a : h'\right]\right)$$

and as we have seen, s_n^h is coherent:

$$i_n^h\left(h'\right) = b_n^h\left(h'\right).$$

By taking the limit along the subsequence $(n_k)_{k\in\mathbb{N}}$ and using continuity, we obtain:

$$b^h\left(h'\right) = \hat{a}^{h'} \in \underset{a\in A_{\omega\left(h'\right)}}{\operatorname{argmax}} U_b^{h'}\left(s^h\left[a : h'\right]\right)$$

and

$$i^h\left(h'\right) = b^h\left(h'\right).$$

Thus, f is a consistent sophisticatedly optimal frame indeed. □

Again, we might ask whether there is a unique consistent sophisticatedly optimal frame. It turns out that this is not the case—it is possible to construct decision problems with multiple sophisticatedly optimal intention–belief pairs for each history. Also, the induced utilities of optimal frames can differ. Thus, only in some cases do we get a unique sophisticatedly optimal frame, and with it, unique predictions for a sophisticated decision maker's intention, beliefs, and induced utility.[19] For these cases Theorem 4 implies that the frame will be consistent, too.

When allowing only for pure actions, a sophisticatedly optimal frame of stationary intention–belief pairs might not exist for some, relatively simple decision problems. While it is natural to associate naiveté with the lack of stationarity ("I will smoke today, but I will quit from tomorrow morning"), the finding that sophisticates might need to resort to non-stationary intention–belief pairs might be somewhat surprising. The implication of this is that it can be too restrictive to focus only on stationary intention–belief pairs.[20]

Overall, our understanding of sophistication in terms of sophisticated selves first working through their beliefs, then deriving their intentions is closest to the self-awareness interpretation. However, sophisticated optimality can indeed be regarded as a notion of intra-personal subgame-perfection.[21] As there might be several such equilibria, selves at various histories might pick actions corresponding to different equilibria; thus, there is no a priori guarantee for the satisfaction of rational expectations, or that the subgame-perfect equilibrium chosen by the root self will actually be followed through. Thus, the equilibrium aspect of sophisticated optimality on the level of *intention–belief pairs* does not imply consistency or stationarity for the sophisticated *frame*. However, the above theorem shows that a consistent sophisticatedly optimal frame exists.

7. Hybrid decision makers

This section introduces a new type of decision maker. So far, we have only considered decision makers who are either always naive, or always sophisticated. However, such purity is quite rare, perhaps even non-existent in the real world. Even the most naive individual realizes after a while that her intentions might not be credible; and even the most consistently sophisticated individual can slip into wishful thinking about her future actions. Therefore, we try to model this duality of an individual via hybrid types. In contrast to the partially naifs of O'Donoghue and Rabin (2001) who (in the context of quasi-hyperbolic discounting) are aware of future present-biasedness, but underestimate its magnitude, our hybrid decision maker flip-flops between being sophisticated and naive.[22]

We now extend our model to capture hybrid types. Our type space includes naifs and sophisticates.

Definition 20. A *Markov decision problem with self types* is made up of:

- the set of time periods $\{0, 1, 2, \ldots, o\}$;

[19] On the non-uniqueness of sophisticatedly optimal strategies, see Phelps and Pollak (1968), Peleg and Yaari (1973) and Blackorby et al. (1973). More recently, Vieille and Weibull (2009) show that non-uniqueness is a generic property for hyperbolic discounting, and also give sufficient conditions for uniqueness. For a refinement concept, see Kocherlakota (1996).

[20] This also implies that good advice need not take the law or rulelike form of "When in X, always do Y!", but instead needs to give leeway to either history, or timing ("When in X, and after having done Z, do Y!" or "When in X on Saturdays, do Y!").

[21] This also highlights why in Definition 5 we defined intention–belief pairs for *all* future histories, since – as in subgame-perfection – the actions of selves at histories off the optimal path are relevant.

[22] In fact, the classical example of Ulysses and the sirens (Elster, 1979) can be reinterpreted along similar lines: Ulysses listening to the sirens would not lose all agency. He would instead (naively and falsely) believe that he can get somewhat closer to the island of the sirens, and stop at a certain distance. The (sophisticated) Ulysses that contemplates this situation in advance realizes the song of the sirens will turn him naive, and ties himself to the mast, preventing his future naive self from making any choice at all. In contrast, in the standard interpretation, Ulysses listening to the sirens is not merely naive, but incapable of rational thought, and thus, choice.

- $\Theta \subseteq \Omega \times X$, where Ω is a finite state space and $X = \{N, S\}$ is the finite type space[23]; we denote a state-type pair by θ; the initial state-type pair is $\theta_0 \in \Theta$; the state component is denoted by $\omega(\theta)$; while the type component is denoted by $x(\theta)$;
- a finite and nonempty set of pure actions A_ω that the decision maker can choose from in ω;
- a payoff function $u_\omega : A_\omega \to \mathbb{R}$ that assigns a payoff to every action in state ω;
- transition probabilities $m_\theta : A_{\omega(\theta)} \to \Delta(\Theta)$, with $m_\theta\left(\theta' | a_{\omega(\theta)}\right)$ denoting the probability to transit from the state-type pair θ to the state-type pair θ' when action $a_{\omega(\theta)}$ is chosen.

This definition keeps the Markovian properties of the original model, and adds a specification of naiveté or sophistication to each state. Also, Θ is common knowledge among the selves.

Our model is quite general. Before moving on to illustrate its use in detail for our motivating example from Section 1, we list a few kinds of decision problems for which it could be used:

- exogenous types: All state-type combinations are allowed ($\Theta = \Omega \times X$). Moreover, in each new state, the self is naive (or sophisticated) with the same probability: $m_\theta((\omega', x)|a_{\omega(\theta)}) = m_\theta((\omega'', x)|a'_{\omega(\theta)})$ for all $\theta, \omega', \omega'', x, a_{\omega(\theta)}, a'_{\omega(\theta)}$. Such a model assumes no correlation between state and type.
- fixed type for each state: Here, we have $(\omega, x), (\omega, x') \in \Theta \Rightarrow x = x'$. This is the opposite of the previous scenario, as there is perfect correlation between state and type. It can be easily seen that any Markov decision problem with self types can be reformulated in this manner by expanding the state space; however, it can make the model less illuminating, possibly hiding structural similarities between the problems faced by a naive and a sophisticated self.
- deterministic type determination: This requires that for all $\theta, a_{\omega(\theta)}$, there is some $x \in X$, such that $\sum_{\omega'} m_\theta((\omega', x)|a_{\omega(\theta)}) = 1$. This means that whatever the self chooses, her type (but not necessarily her state) in the next period is fully determined.
- full control over type: For all $\theta, x \in X$, there is some $a_{\omega(\theta)}$ so that $\sum_{\omega'} m_\theta((\omega', x)|a_{\omega(\theta)}) = 1$. Each self can always determine the type of the next period self.

Of course, these are boundary cases, and many interesting situations lie in the middle, having some, but imperfect correlation between state and type; and giving some, but less than total control for selves over their future types. Indeed, it could be argued that the drinking problem we present below should involve stochastic rather than deterministic type determination. However, for simplicity of analysis, we abstract from such complications.

To understand optimal intention–belief pairs for hybrid selves, we first have to re-define the notion of history:

Definition 21. A *type-dependent history* h has the form $h = \left(\theta_0, a_{\omega(\theta_0)}, \ldots, \theta_{t-1}, a_{\omega(\theta_{t-1})}, \theta_t\right)$, with:

- $\theta_i \in \Theta$, for $i \in \{0, 1, \ldots, t\}$;
- $a_{\omega(\theta_i)} \in A_{\omega_i}$, for $i \in \{0, 1, \ldots, t-1\}$;
- $m_{\theta_i}\left(\theta_{i+1} | a_{\omega(\theta_i)}\right) > 0$, for $i \in \{0, 1, \ldots, t-1\}$.

Extending the previous notation, $x(h)$ refers to the current type. We keep the association between histories and selves—each history now corresponds to a self, and it also includes the self's type.

Next, we define optimal type-dependent intention–belief pairs for the Markov decision problem with self types. We make use of Definitions 18 and 19 for naively and sophisticatedly optimal intention–belief pairs. We first present the formal definition, and then explain the intuitions below.

Definition 22. A type-dependent intention–belief pair s^h for a Markov decision problem with self types is *optimal* at history h, if it satisfies the following conditions:

- for $x(h) = N$:

  $$s^h \in \operatorname*{argmax}_{s \in S^h} U_i^h(s),$$

  and

  $$b^h\left(h'\right) = i^h\left(h'\right), \quad \text{for all } h' \in H^{\triangleright h}$$

- for $x(h) = S$:

  $$b^h\left(h'\right) \in \left[\operatorname*{argmax}_{s \in S^{h'}} U_i^{h'}(s)\right]\left(h'\right),$$

  for all $h' \in H^{\triangleright h}$ with $x\left(h'\right) = N$; [24]

  $$b^h\left(h'\right) \in \operatorname*{argmax}_{a \in A_{\omega\left(h'\right)}} U_b^{h'}\left(s^{h'}\left[a : h'\right]\right),$$

  for all $h' \in H^{\triangleright h}$ with $x\left(h'\right) = S$;

  and

  $$i^h\left(h'\right) \in \operatorname*{argmax}_{a \in A_{\omega\left(h'\right)}} U_b^{h'}\left(s^{h'}\left[a : h'\right]\right),$$

  for all $h' \in H^{\triangleright h}$ with $x\left(h'\right) = N$;

  $$i^h\left(h'\right) = b^h\left(h'\right), \quad \text{for all } h' \in H^{\triangleright h} \text{ with } x\left(h'\right) = S.$$

A type-dependent frame f is optimal, if $f(h)$ is an optimal type-dependent intention–belief pair for all h.

Although this definition is rather lengthy, it captures our basic intuitions for the two types. A naive self at h does not reason about future selves, as her intentions determine her beliefs, and as a result, her whole optimal intention–belief pair in the standard way. However, a sophisticated self at h is able to reason about future selves in the following manner: If a future self at h' is naive, then the self at h believes the self at h' will act in a naive way, maximizing her expected utility based on intentions. If, on the other hand, a future self at h' is sophisticated, then the self at h (correctly) believes that the self at h' will act in a sophisticated way, being able to reason about future selves just as well as h herself does. So a sophisticated self at h intends to choose in a sophisticated manner at all nodes, giving a best response to the choices of future selves. This implies that the intention and belief component of an optimal type-dependent intention–belief pair match for all future histories where the self is sophisticated, but they might not match with those of future naive selves. Coherence is therefore not a necessary property of optimal type-dependent intention–belief pairs.

Naively optimal intention–belief pairs are, in general, not stationary, and thus the belief component of an optimal type-dependent frame can also be non-stationary. Moreover, it is easy to see that such a frame is not consistent: the intentions of sophisticated selves will not, in general, correspond to the actions taken by naive selves. In both examples of hybrid decision making we present below, we will see such inconsistencies.[25] The lack of coherence, stationarity and consistency represent the inner conflicts that arise within a hybrid decision maker.

[23] N stands for naive, and S for sophisticated.

[24] Technically, $\left[\operatorname{argmax}_{s \in S^{h'}} U_i^{h'}(s)\right]\left(h'\right)$ is a pair of actions (an intention–belief pair), whereas $b^h\left(h'\right)$ is only one action. However, due to Definition 5 – requiring that the intentions and beliefs of a self at h' for the action at h' coincide – it does not matter which element of the pair is taken.

[25] Recall that when an action is actually taken by a self, it is both believed and intended by the self for the current history.

Theorem 5. *For any decision problem, there exists a type-dependent optimal frame.*

Proof. First, start with determining the intentions and beliefs of naive selves, i.e., those at histories where $x(h) = N$. For these, we can simply use the first part of the proof of Theorem 3. So, we have $s^h = (i^h, b^h)$ defined for all h with $x(h) = N$. Let $f(h) = s^h$ for all such h.

Moving now to histories at which the self is sophisticated, our construction is analogous to that of Theorem 4. Let $\mathscr{A} = \times_{h\in H} A_{\omega(h)}$, and take an arbitrary $a = (a^h)_{h\in H} \in \mathscr{A}$ that assigns an action to each history.

Recall that a type-dependent optimal frame will not necessarily be coherent. Therefore, both intentions and beliefs have to be constructed; we start with beliefs. First, transform a into $\check{a}$ by fixing actions assigned to histories where the self is naive, i.e., where $x(h) = N$, so that they are actions corresponding to those selves acting in a naive way:

$$\check{a}^h = i^h(h), \quad \text{if } x(h) = N;$$
$$\check{a}^h = a^h, \quad \text{if } x(h) = S.$$

Let $a_n \in \mathscr{A}$ be defined as follows: $a_n^h = \check{a}^h$ for all h with $t(h) > n$, or with $t(h) \leq n$, and $x(h) = N$. For the remaining histories with $t(h) \leq N$ and $x(h) = S$, we move by backwards induction from n to 0, and let a_n^h be the best response to the future actions, which are already all defined.

Thus, we obtain a sequence $(a_n)_{n\in\mathbb{N}}$ in $\mathscr{A}$. Since $\mathscr{A}$ is sequentially compact (see the proof of Theorem 4), the sequence a_n has a subsequence a_{n_k} converging to some $\hat{a} \in \mathscr{A}$.[26] Thus, for all possible choices of a horizon m, we can find a K_m such that $\hat{a}^h = a_{n_k}^h$ for all $k \geq K_m$ and h with $t(h) \leq m$.

Now, let the beliefs of a sophisticated self – i.e., at a history h with $x(h) = S$ – be $b^h(h') = \hat{a}^{h'}$.

Finally, we construct the intentions of sophisticated selves. Set sophisticated selves' beliefs about future nodes as: $i^h(h') = b^h(h')$ for all h' with $x(h) = x(h')$. For sophisticated selves' intentions assigned to future naive nodes, we construct $\bar{a}$ by modifying $\hat{a}$ in a way that whenever $x(h) = N$, $\bar{a}^h$ is a best response to the future actions in $\hat{a}$. We keep actions at other, sophisticated histories unchanged: $\bar{a} = \hat{a}$. We set the intentions of a sophisticated self to be $i^h(h') = \bar{a}^{h'}$.

We have thus defined both b^h and i^h for sophisticated selves. Let $s^h = (i^h, b^h)$ and finally, $f(h) = s^h$ for all $x(h) = S$. This completes our construction of a type-dependent optimal frame, as we have provided the intention–belief pairs assigned to histories where the self is naive, and also to the ones where she is sophisticated.

We will now review our construction again to confirm that f is indeed a type-dependent optimal frame. For $x(h) = N$, this is immediate. For $x(h) = S$, we will first check the beliefs, and then the intentions.

First, suppose that $x(h') = N$. We have:

$$b^h(h') = \hat{a}^{h'} = i^{h'}(h') \in \operatorname*{argmax}_{s\in S^{h'}} U_i^{h'}(s)(h'),$$

which is what is required in the definition of f. For the other case, take $x(h') = S$. Now, from our construction of $a_n^{h'}$, for all $n \geq t(h')$:

$$b_n^h(h') = a_n^{h'} \in \operatorname*{argmax}_{a\in A_{\omega(h')}} U_b^{h'}(b_n^h[a : h']).$$

[26] There might be multiple subsequences converging to different $\hat{a}$-s; in that case, we can select any one of them.

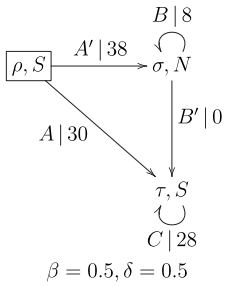

Fig. 7. A drinking problem.

Taking the limit along the subsequence n_k, using continuity, we get:

$$b^h(h') = a^{h'} \in \operatorname*{argmax}_{a\in A_{\omega(h')}} U_b^{h'}(b^h[a : h']).$$

For the intentions of sophisticated selves for future sophisticated nodes, we set these directly to be $i^h(h') = b^h(h')$. The last thing we need to check is the intentions of sophisticated selves for future naive selves. These were defined as:

$$i^h(h') = \bar{a}^{h'} \in \operatorname*{argmax}_{a\in A_{\omega(h')}} U_b^{h'}(b^{h'}[a : h']). \quad \square$$

We now return to the problem raised in the introduction, displayed in Fig. 7. The decision maker is sitting in a pub, having finished her second beer (in state ρ), and is of type S (sophisticated). She can either go home directly by choosing action A, transiting to state τ, where she does not need to make any more decisions. Alternatively, she can drink 'one more beer' by choosing action B. This, however, transitions her to a 'drunken' state σ, where she becomes type N (naive). In this drunken state, she can choose between drinking one more beer by choosing B (thus, maintaining her drunkenness and staying at σ), or going home to state τ by choosing B'.

Let $h_0 = ((\rho, S))$ be the root history. Our goal is to construct s^{h_0}, the optimal type-dependent intention–belief pair for the root self at this history. Since $x(h_0) = S$, we have to deal with the interesting case, that of a sophisticated root self.

We start the analysis of this situation by focusing on the state σ. Take any h' such that $\omega(h') = \sigma$. As $x\left(h'\right) = N$ if $\omega\left(h'\right) = \sigma$, we get $b^h\left(h'\right) \in \operatorname{argmax}_{s\in S^{h'}} U_i^{h'}(s)\left(h'\right)$ according to the definition; the root self believes that a self at history h' with a current state σ is naive. What is the naive choice in state σ? It is relatively easy to see that the (subgame-optimal) naively optimal intention–belief pair is $(B, B)(B', B')(B', B')\ldots$, i.e., 'drink one more beer, and then go home'. Thus, $b^{h_0}(h') = B$ whenever $\omega(h') = \sigma$. The root self thus believes she would continue drinking after becoming naive.

What about the intentions of the root self for the self at σ? She believes that at history h', the continuation actions will be B. The self at history h' could only choose between B and B'. Going for B yields $8 + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{8}{1-\frac{1}{2}} = 12$, whereas picking B' gives $0 + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{28}{1-\frac{1}{2}} = 14$. Therefore, $i^{h_0}\left(h'\right) = B'$ whenever $\omega(h') = \sigma$. Together with the result of the previous paragraph, we get that $s^{h_0}\left(h'\right) = (B', B)$ whenever $\omega(h') = \sigma$. We see that the intentions and the beliefs of the root self do not match: She would like future selves to pick B', but correctly anticipates that future selves will be unable to do so, and would actually choose B. The sober, sophisticated root self realizes that if she drinks just one more beer, she will end up drinking much more than what she actually wishes for.

So what should the root self choose at h_0? She can pick A, going home directly, earning her $U_b^{h_0}(s^{h_0}[A : h_0]) = 30 + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{28}{1-\frac{1}{2}} = 44$. Or, she can pick A', drink one more beer, and end up still being in

the pub. This would earn her $U_b^{h_0}(s^{h_0}[A' : h_0]) = 38 + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{8}{1-\frac{1}{2}} = 42$. Going home seems best. Thus, $s^{h_0}(h_0) = (A, A)$. The optimal type-dependent intention–belief pair is:

$$
\begin{aligned}
&s^{h_0}(h_0) = (A, A),\\
&s^{h_0}(h') = (B', B), \quad \text{whenever } \omega(h') = \sigma,\\
&s^{h_0}(h') = (C, C), \quad \text{whenever } \omega(h') = \tau.
\end{aligned}
$$

Note that a fully naive root self would expect that she can resist the temptation of drinking additional bottles of beer, and would expect a utility of $38 + \frac{1}{2} \cdot \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2}^2 \cdot \frac{28}{1-\frac{1}{2}} = 45$. The sophisticated root self realizes that this is unattainable, as the incentives *and the type* of the selves change by transiting to σ. So in the drinking problem, the sober, sophisticated root self avoids becoming naive, and thus is better off. Sophistication thus can help avoiding the trap of naiveté. But can sophistication help in avoiding the pitfalls of sophistication?

O'Donoghue and Rabin (1999a) have shown that in some decision problems, a naive decision maker is strictly better off than a sophisticated one. The psychological intuition behind this is that being aware of one's inability to resist a temptation in the future can render one unable to resist the temptation in the present, too. This observation allows us to construct the decision problem in Fig. 8, which we call the indulgence problem with hybrid type.

The following is an example of the indulgence problem: When to consume a bottle of valuable wine that gains in quality for up to three years, but then becomes undrinkable? Should one consume it immediately (B'), or after one, two, or three years (C', D', or E')? The problem in Fig. 8 complicates this by adding an initial choice of type: The (sophisticated) root self needs to decide whether to face the indulgence problem while being *sophisticated* or *naive*.

Since she is sophisticated, the root self at history $h_0 = ((\rho, S))$ is able to reason about future selves in the following way:

- At history $h_1 = ((\rho, S), A, (\sigma, S))$, and at all succeeding histories, she will be sophisticated. Reasoning by backwards induction:
  - At (σ'', S),[27] the sophisticated self will decide to consume (D') rather than wait further (D), because $40 > 36 = \frac{1}{2} \cdot \frac{1}{2} \cdot 144$.
  - Thus, the sophisticated self at (σ', S) will also decide to consume (C') rather than wait (C), because she knows that the wine will be consumed in the next period, and $12 > 10 = \frac{1}{2} \cdot \frac{1}{2} \cdot 40$.
  - Along similar lines, at (σ, S) the sophisticated self will again decide to consume (B') rather than wait (B), because she knows that the wine will be consumed in the next period, and $4 > 3 = \frac{1}{2} \cdot \frac{1}{2} \cdot 12$.
  - Thus, if the root self decides to face the indulgence problem while being sophisticated (A), she can expect that the wine will be consumed immediately afterwards (B').
- On the other hand, at history $h_1' = ((\rho, S), A', (\sigma, N))$, and at all succeeding histories, she will be naive. We solve the problem of the naive selves one by one.
  - The naive self at (σ, N) will plan to wait three years, expecting a utility of $9 = \left(\frac{1}{2}\right)^4 \cdot 144$, which is more than waiting two years $(5 = \left(\frac{1}{2}\right)^3 \cdot 40)$, one year $(3 = \left(\frac{1}{2}\right)^2 \cdot 12)$, or consuming immediately (4). Therefore, the naive self at (σ, N) will choose to wait (B).

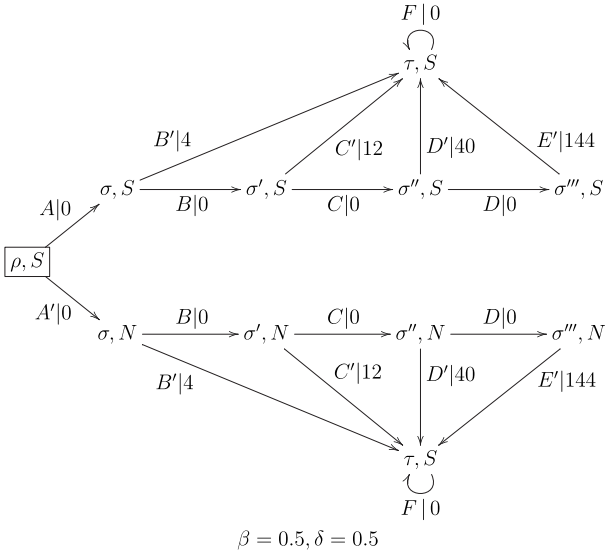

Fig. 8. Optimal self-deception in the indulgence problem with hybrid type.

  - The naive self at (σ', N) will plan to wait two more years, expecting a utility of $18 = \left(\frac{1}{2}\right)^3 \cdot 144$, which is more than waiting one more year $(10 = \left(\frac{1}{2}\right)^2 \cdot 40)$, or consuming immediately (12). Therefore, the naive self at (σ', N) will also choose to wait (C).
  - The naive self at (σ'', N) will indulge and drink the wine immediately (D'), gaining 40, which is more than what she can gain by waiting $(D$, which yields $36 = \left(\frac{1}{2}\right)^2 \cdot 144)$.
  - Thus, if the root self decides to face the indulgence problem while being naive (A'), she can expect that the wine will be consumed after two years (D').

Now, from the perspective of the root self:

$$
\begin{aligned}
U_b^{h_0}(s^{h_0}[A : h_0]) &= \frac{1}{2} \cdot \frac{1}{2} \cdot 4 = 1 < 2.5 = \frac{1}{2} \cdot \left(\frac{1}{2}\right)^3 \cdot 40\\
&= U_b^{h_0}(s^{h_0}[A' : h_0]).
\end{aligned}
$$

Therefore, the best response of the root self is to choose action A'. Notice that this means that a sophisticated root self chooses to face the indulgence problem as a naif. Thereby the self at h_0 intentionally causes the self at h_1' to have wrong beliefs. In particular, the naive self at h_1' believes that she will be able to wait until the wine fully matures, and takes action E'. The sophisticated self at the preceding history h_0 knows that this is not the case, that in fact, action D' will be taken. Thus, when choosing an optimal type-dependent intention–belief pair, the root self realizes that she is better off with false beliefs, and decides to deceive herself.

By what means such self-deception might be effectively achieved, or whether it can be achieved intentionally at all is, of course, a difficult problem. But it seems like self-deception has its virtues, which might, in itself, challenge ethical arguments on the inherent immorality of self-deception.[28] Optimal self-deception strengthens previous results showing that a decision maker that cannot commit her consumption plan might optimally avoid acquiring information, even when learning that information would be free (Carrillo and Mariotti, 2000).

[27] Since in this decision problem there is a single history reaching each state-type pair, we will talk about the selves being at a state-type pair, instead of writing out the full history.

[28] For an overview on the philosophical problems of self-deception, see Deweese-Boyd (2012).

8. Concluding remarks and future research

This work attempts to play a foundational role for future discourse in multi-self models of dynamic inconsistency. It establishes that the basic epistemic concepts to be considered are beliefs and intentions, and the main levels of analysis should be those of intention–belief pairs and frames. We would now like to provide some remarks and outline some directions for follow-up research in this area.

An obvious limitation of the current framework is that it only allows for pure actions. This limitation is introduced to ease the presentation, but the technical adaptations required for dealing with mixed actions can be accomplished rather straightforwardly. Mixed actions should play a particularly important role when moving from decision-theoretic models to a game setting.

One might wonder how flexible this model is with regards to increasing the state or action space of the decision problem. Countably infinite states and actions can be allowed for without much difficulty as long as the set of payoffs for each action remains compact (and hence, bounded). However, handling continuous time would require a fundamentally different framework, along with a reinterpretation of the notion of 'self'.

Whereas our focus was the two most common types of decision makers facing dynamic inconsistency, naifs and sophisticates, there have been arguments in the literature for taking into account other types as well. In particular, McClennen (1990) argues for the possibility of resolute decision making. Actually, resoluteness can easily be incorporated into our framework. Expand the type space in Markov decision problems with hybrid types to include resolutes could allow for modeling an even broader class of phenomena.

The horizon of sophisticated decision makers requires further investigation. If selves possess only a finite horizon, reasoning about future selves can be based on two assumptions: Either the *length*, or the *endpoint* of the horizon of that future self is the same as that of the current self. In the former case, we are talking about a *moving*, in the latter, about a *fixed* horizon. The implications of these two assumptions on optimal decisions (derived, for instance, via backward induction) are not yet understood.

Finally, the most interesting application of the framework presented above will be for game theory. How can players reason about the intentions and beliefs of other players, as well as their types? How can one exploit the naiveté (or sophistication) of others? What kind of equilibria are generated when naive, sophisticated, or hybrid players are pitted against each other? We hope that through this work, we have broken the ground for such questions.

Acknowledgments

The authors wish to thank Jean-Jacques Herings and four anonymous reviewers for their invaluable comments. The Netherlands Organisation for Scientific Research (NWO; 452-08-006) is gratefully acknowledged for its support.

References

Akerlof, G.A., 1991. Procrastination and obedience. Amer. Econ. Rev. 81 (2), 1–19.
Ali, S.N., 2011. Learning self-control. Quart. J. Econ. 126 (2), 857–893.
Angeletos, G.M., Laibson, D.I., Repetto, A., Tobacman, J., Weinberg, S., 2001. The hyperbolic consumption model: Calibration, simulation, and empirical evaluation. J. Econ. Perspect. 15 (3), 47–68.
Asheim, G.B., 2007. Procrastination, partial naivete, and behavioral welfare analysis. No. 2007, 02. Memorandum, Department of Economics, University of Oslo.
Bach, C.W., Heilmann, C., 2011. Agent connectedness and backward induction. Int. Game Theory Rev. 13 (02), 195–208.
Benabou, R., Pycia, M., 2002. Dynamic inconsistency and self-control: A planner-doer interpretation. Econom. Lett. 77 (3), 419–424.
Bernheim, B.D., Rangel, A., 2004. Addiction and cue-triggered decision processes. Amer. Econ. Rev. 94 (5), 1558–1590.
Blackorby, C., Nissen, D., Primont, D., Russell, R.R., 1973. Consistent intertemporal decision making. Rev. Econom. Stud. 40 (2), 239–248.
Brocas, I., Carrillo, J.D., 2008. The brain as a hierarchical organization. Amer. Econ. Rev. 98 (4), 1312–1346.
Carrillo, J.D., Mariotti, T., 2000. Strategic ignorance as a self-disciplining device. Rev. Econom. Stud. 67 (3), 529–544.
Cowen, T., 1991. Self-constraint versus self-liberation. Ethics 101 (2), 360–373.
DellaVigna, S., Malmendier, U., 2004. Contract design and self-control: Theory and evidence. Quart. J. Econ. 119 (2), 353–402.
DellaVigna, S., Malmendier, U., 2006. Paying not to go to the gym. Amer. Econ. Rev. 694–719.
Deweese-Boyd, I., 2012. Self-deception. The Stanford Encyclopedia of Philosophy (Spring 2012 Edition), Edward N. Zalta (ed.), Retrieved from http://plato.stanford.edu/archives/spr2012/entries/self-deception.
Eliaz, K., Spiegler, R., 2006. Contracting with diversely naive agents. Rev. Econom. Stud. 73 (3), 689–714.
Elster, J., 1979. Ulysses and the Sirens: Studies in Rationality and Irrationality. Cambridge University Press, Cambridge.
Fischer, C., 1999. Read this paper even later: Procrastination with time-inconsistent preferences. Discussion Paper 99–20, Resources for the Future.
Fudenberg, D., Levine, D.K., 1983. Subgame-perfect equilibria of finite- and infinite-horizon games. J. Econom. Theory 31 (2), 251–268.
Fudenberg, D., Levine, D.K., 2006. A dual-self model of impulse control. Amer. Econ. Rev. 96 (5), 1449–1476.
Fudenberg, D., Levine, D.K., 2012. Timing and self-control. Econometrica 80 (1), 1–42.
Gilpatric, S.M., 2008. Present-biased preferences, self-awareness and shirking. J. Econ. Behav. Organ. 67 (3), 735–754.
Gruber, J., Kőszegi, B., 2001. Is addiction rational? Theory and evidence. Quart. J. Econ. 116 (4), 1261–1303.
Gul, F., Pesendorfer, W., 2001. Temptation and self-control. Econometrica 69 (6), 1403–1435.
Hammond, P.J., 1976. Changing tastes and coherent dynamic choice. Rev. Econom. Stud. 43 (1), 159–173.
Harris, C., Laibson, D.I., 2001. Dynamic choices of hyperbolic consumers. Econometrica 69 (4), 935–957.
Heidhues, P., Kőszegi, B., 2009. Futile attempts at self-control. JEEA 7 (2–3), 423–434.
Herings, P.J.J., Rohde, K.I., 2006. Time-inconsistent preferences in a general equilibrium model. Econom. Theory 29 (3), 591–619.
Herings, P.J.J., Rohde, K.I., 2008. On the completeness of complete markets. Econom. Theory 37 (2), 171–201.
Jehiel, P., Lilico, A., 2010. Smoking today and stopping tomorrow: A limited foresight perspective. CESifo Econ. Stud. 56 (2), 141–164.
Kahneman, D., Wakker, P.P., Sarin, R., 1997. Back to Bentham? Explorations of experienced utility. Quart. J. Econ. 112 (2), 375–406.
Kavka, G.S., 1983. The toxin puzzle. Analysis 43 (1), 33–36.
Kocherlakota, N.R., 1996. Reconsideration-proofness: A refinement for infinite horizon time inconsistency. Games Econom. Behav. 15 (1), 33–54.
Laibson, D.I., 1994. Hyperbolic discounting and consumption (Doctoral dissertation), Massachusetts Institute of Technology.
Laibson, D.I., 1997. Golden eggs and hyperbolic discounting. Quart. J. Econ. 112 (2), 443–478.
Lapied, A., Renault, O., 2012. A subjective discounted utility model. Econ. Bull. 32 (2), 1171–1179.
Loewenstein, G., 2005. Hot-cold empathy gaps and medical decision making. Health Psychol. 24 (4S), S49–S56.
Maskin, E., Tirole, J., 2001. Markov perfect equilibrium: I. Observable actions. J. Econom. Theory 100 (2), 191–219.
McClennen, E.F., 1990. Rationality and Dynamic Choice. Cambridge University Press, Cambridge.
O'Donoghue, T., Rabin, M., 1999a. Doing it now or later. Amer. Econ. Rev. 89 (1), 103–124.
O'Donoghue, T., Rabin, M., 1999b. Incentives for procrastinators. Quart. J. Econ. 114 (3), 769–816.
O'Donoghue, T., Rabin, M., 2001. Choice and procrastination. Quart. J. Econ. 16 (1), 121–160.
Peleg, B., Yaari, M.E., 1973. On the existence of a consistent course of action when tastes are changing. Rev. Econom. Stud. 40 (3), 391–401.
Phelps, E.S., Pollak, R.A., 1968. On second-best national saving and game-equilibrium growth. Rev. Econom. Stud. 35 (2), 185–199.
Read, D., Frederick, S., Orsel, B., Rahman, J., 2005. Four score and seven years from now: The date/delay effect in temporal discounting. Manage. Sci. 51 (9), 1326–1335.
Sáez-Martí, M., Weibull, J.W., 2005. Discounting and altruism to future decision-makers. J. Econom. Theory 122 (2), 254–266.

Sarafidis, Y., 2004. Inter-temporal price discrimination with time inconsistent consumers. In: Econometric Society 2004 North American Summer Meetings (No. 479). Econometric Society.
Strotz, R.H., 1956. Myopia and inconsistency in dynamic utility maximization. Rev. Econom. Stud. 23 (3), 165–180.
Thaler, R.H., Shefrin, H.M., 1981. An economic theory of self-control. J. Polit. Econ. 89 (2), 392–406.
Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211 (4481), 453–458.
Vieille, N., Weibull, J.W., 2009. Multiple solutions under quasi-exponential discounting. Econom. Theory 39 (3), 513–526.
Yaari, M.E., 1978. Endogenous changes in tastes: A philosophical discussion. In: Eberlein, G., Leinfellner, W. (Eds.), Decision Theory and Social Ethics. Springer Netherlands, pp. 59–98.